We present the syntactically supervised Transformer (SynST), which achieves faster translation and higher BLEU than competing non-autoregressive neural machine translation models.

**SYNST VS. EXISTING SYSTEMS**

*Gold Parse*  
**Predicted Parse:**  
*Source:* Only predicting constituent length (\(k = 1\))  
*Constituent identity is crucial for quality*

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU 2018</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla Transformer</td>
<td>22.78</td>
<td>1.00</td>
</tr>
</tbody>
</table>
| Semi-Autoregressive Transformer  
\(k = 2\) | 22.81 | 2.05 |
\(k = 4\) | 16.44 | 3.61 |
\(k = 6\) | 12.55 | 4.86 |
| Latent Transformer* | 22.81 | 2.05 |
*As reported in (Kaiser et al. 2018)*

**SynST:**  
\(NP1 > VP3\) > Cats sleep a lot  
During inference, the model uses its own chunk predictions.

**TARGET PARSE CHUNKING**

During an in-order traversal, if the subtree rooted at a visited node spans \(k\) tokens, append it to our **chunk sequence**.

**ANALYSIS ON IWSLT DEV SET**

Constituent identity is crucial for quality

*Only predicting constituent length (\(1 > 3\)) rather than type & length (\(NP1 > VP3\)), causes a BLEU drop from 23.8 to 8.2.*

**Ground-truth syntax yields huge improvements**

Conditioning on the ground-truth chunk sequence during inference dramatically **improves BLEU** from 23.8 to 41.5, yielding an upper bound for our approach.

**How much does SynST rely on syntax?**

**Source:** Katzen schlafen viel  
**Predicted Parse:** NP1 > VP2  
**Gold Parse:** NP1 > VP3

**Target:** Cats sleep a lot  
**Prediction:** Cats sleep lots  
**Parsed Prediction:** NP1 > VP2

**Future work:**

**dynamic vs fixed \(k\)**

Randomly sampling possible chunk sequences during training by varying \(k\) leads to a **large BLEU improvement** (+1.5) with minimal impact to speedup (drop from 3.8x to 3.1x). Improving parse prediction is an avenue for future research.

**SynST’s bottleneck is its parse decoder**

A one-layer parse decoder is ~3x faster than a 5-layer version, with only a ~0.5 BLEU drop.

**CONTROLLED EXPERIMENTS**

**Model**  
**WMT En–De**  
**BLEU**  
**Speedup**  
**WMT De–En**  
**BLEU**  
**Speedup**  
**IWSLT En–De**  
**BLEU**  
**Speedup**  
**WMT En–Fr**  
**BLEU**  
**Speedup**

<table>
<thead>
<tr>
<th>Model</th>
<th>WMT En–De</th>
<th>WMT De–En</th>
<th>IWSLT En–De</th>
<th>WMT En–Fr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla Transformer</td>
<td>26.87</td>
<td>30.73</td>
<td>30.00</td>
<td>40.22</td>
</tr>
<tr>
<td>Beam Size = 4</td>
<td>1.00x</td>
<td>1.00x</td>
<td>1.00x</td>
<td>1.00x</td>
</tr>
<tr>
<td>Beam Size = 1</td>
<td>25.82</td>
<td>29.63</td>
<td>28.66</td>
<td>39.41</td>
</tr>
</tbody>
</table>
| Semi-Autoregressive Transformer  
\(k = 2\) | 22.81 | 26.78 | 25.48 | 36.62 |
\(k = 4\) | 16.44 | 21.27 | 20.25 | 28.07 |
\(k = 6\) | 12.55 | 15.23 | 14.02 | 24.63 |
| Latent Transformer*    | 19.8     | 3.89x     | 14.02       | -         |
| *As reported in (Kaiser et al. 2018)* |
| Syntactically Supervised Transformer  
\(k = 6\) | 20.74 | 4.86x | 23.82 | 33.47 |
|                       |           |           | 5.32x       |           |